

WHEN ARE THE MOST INFORMATIVE COMPONENTS FOR INFERENCE ALSO THE PRINCIPAL COMPONENTS?

RAJ RAO NADAKUDITI

ABSTRACT. Which components of the singular value decomposition of a signal-plus-noise data matrix are most informative for the inferential task of detecting or estimating an embedded low-rank signal matrix? Principal component analysis ascribes greater importance to the components that capture the greatest variation, *i.e.*, the singular vectors associated with the largest singular values. This choice is often justified by invoking the Eckart-Young theorem even though that work addresses the problem of how to best *represent* a signal-plus-noise matrix using a low-rank approximation and *not* how to best *infer* the underlying low-rank signal component.

Here we take a first-principles approach in which we start with a signal-plus-noise data matrix and show how the spectrum of the noise-only component governs whether the principal or the middle components of the singular value decomposition of the data matrix will be the informative components for inference.

Simply put, if the noise spectrum is supported on a connected interval, in a sense we make precise, then the use of the principal components is justified. When the noise spectrum is supported on multiple intervals, then the middle components might be more informative than the principal components.

The end result is a proper justification of the use of principal components in the oft considered setting where the noise matrix is i.i.d. Gaussian. An additional consequence of our study is the identification of scenarios, generically involving heterogeneous noise models such as mixtures of Gaussians, where the middle components might be more informative than the principal components so that they may be exploited to extract additional processing gain. In these settings, our results show how the blind use of principal components can lead to suboptimal or even faulty inference because of phase transitions that separate a regime where the principal components are informative from a regime where they are uninformative. We illustrate our findings using numerical simulations and a real-world example.

1. INTRODUCTION

Consider a signal-plus-noise data matrix modeled as

$$\tilde{X} = \sum_{i=1}^r \theta_i u_i v_i^H + X, \quad (1)$$

Date: February 7, 2013.

2000 Mathematics Subject Classification. 15A52, 46L54, 60F99.

Key words and phrases. Random matrices, Haar measure, principal components analysis, informational limit, free probability, phase transition, random eigenvalues, random eigenvectors, random perturbation, sample covariance matrices.

This work was supported by the ARO MURI W911NF-11-1-0391 grant and an AFOSR Young Investigator Award FA9550-12-1-0266. The author thanks Florent Benaych-Georges for many inspiring conversations that led to this work and Iain Johnstone for suggesting Gaussian mixture models as an example where middle components might be informative.

where X denotes the $n \times m$ noise-only matrix and $S = \sum_{i=1}^r \theta_i u_i v_i^H$ is the rank- r signal matrix. Relative to this model, the detection and estimation tasks in signal processing and data analysis deal with inferring the presence of and estimating the rank r matrix S given \tilde{X} .

Principal component analysis plays an important role in the setting where $r \ll \min(m, n)$ as described succinctly by Jolliffe [19, Ch1., pp.1]:

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set ... while *retaining as much as possible of the variation*¹ present in the data set. The ... *first few retain most of the variation*¹ present in all of the original variables.

The *first few* principal components alluded to here refer to the first few singular vectors associated with the largest singular values of \tilde{X} . Working with the hypothesis that the directions of greatest variation of the data set must reflect (or correlate with) the signal content and equipped with the singular value decomposition (SVD) as a technique for computing these directions, we can tackle the detection problem in the following manner.

We start off by computing the SVD of \tilde{X} and plot the singular values $\{\tilde{\sigma}_i\}_{i=1}^n$ in non-increasing order. We then estimate the rank of the latent signal matrix S based on the rule:

$$\hat{r} = \{\text{First } i \text{ such that } \text{gap}(i) := \tilde{\sigma}_i - \sigma_{\text{null}} < \text{threshold}\} - 1, \quad (2)$$

where σ_{null} is the largest singular value of the noise-only matrix X which is assumed (in the simplest setting) to be known. This rule, and other modifications thereof, yields an estimate \hat{r} for the rank of the latent signal matrix; when $\hat{r} > 0$ we have detected a signal matrix; see for example [12, Section 14.5] or [19, Section 6.1.3] for classical approaches and [17, 2, 3, 18, 11, 27, 24, 26, 21, 22, 25] for recent random matrix-theoretic approaches.

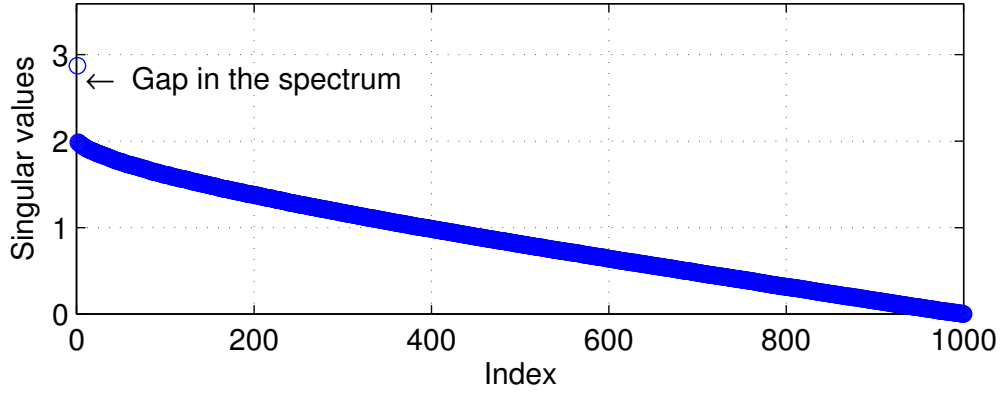
The estimation problem is similarly tackled by computing the truncated SVD of \tilde{X} that employs the \hat{r} (leading or) principal components. This yields a rank \hat{r} estimate of the low-rank signal matrix given by

$$\hat{S} = \sum_{i=1}^{\hat{r}} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^H. \quad (3)$$

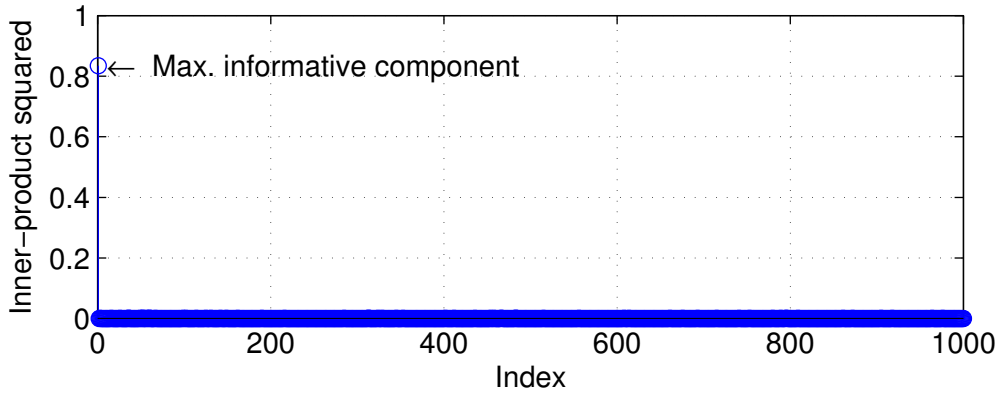
Does the principal component approach to detection and estimation work? Figure 1(a) plots the singular values of a $n \times m$ signal-plus-noise data matrix modeled as $\tilde{X} = 2uv^H + X$, where the noise-only matrix X has i.i.d. mean zero, variance $1/m$ Gaussian entries and the signal matrix $S = 2uv^H$ has rank one. This example, where $n = m = 1000$, illustrates a setting where the gap heuristic in (2) for signal-matrix detection “works” subject to a specification of the gap size threshold.

Figure 1(b) plots the n inner-products $\{|\langle \tilde{u}_i, u \rangle|^2\}_{i=1}^n$, where $\{\tilde{u}_i\}_{i=1}^n$ are the left singular vectors of \tilde{X} . The quantities $\{|\langle \tilde{u}_i, u \rangle|^2\}_{i=1}^n$ (and $\{|\langle \tilde{v}_i, v \rangle|^2\}_{i=1}^m$) are measures of informativeness of the singular vectors of \tilde{X} with respect to the singular vectors of the latent signal matrix. Clearly, the principal left (also, the right - not plotted here) singular

¹Emphasis added.



(a) The singular value spectrum.



(b) PCA example.

FIGURE 1. The singular value spectrum of the signal-plus-noise data matrix exhibits a continuous-looking portion that may be associated with “noise” singular values and a single separated singular value that may be interpreted as evidence of a rank-one “signal” matrix buried in the data matrix.

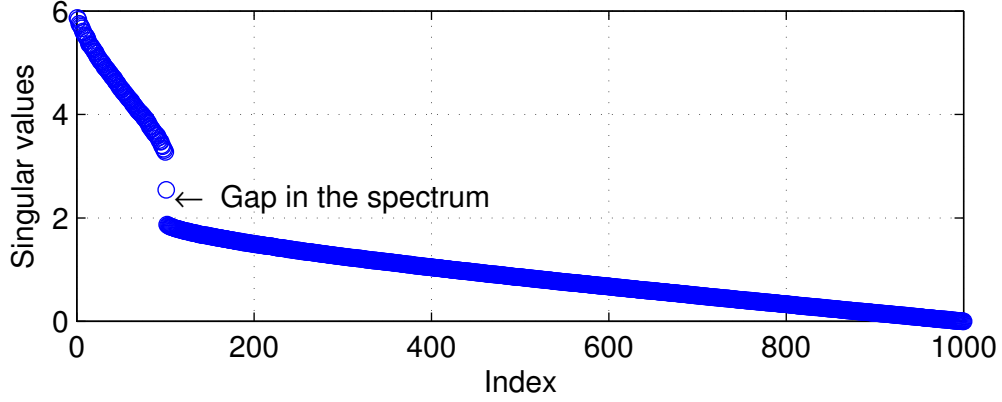
vector is the most informative component and employing it in an estimate of the signal matrix as in (3) is judicious.

Extending the notion of informativeness further, we might define “informative components” as components of the SVD of the data matrix \tilde{X} that are most correlated with the embedded low-rank signal matrix and which consequently best (in a manner to be made precise later) facilitate the detection and estimation tasks described earlier.

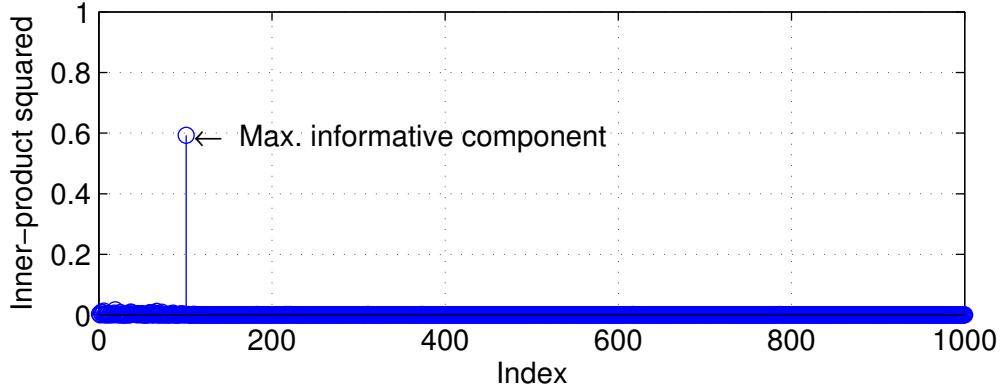
For the example in Figure 1(b), the principal component *is* the most informative component. In other words, the principal component which captures the greatest *variation* in the data is also the component most correlated with the underlying signal matrix. A natural question arises:

Are the most *informative components* necessarily the *principal components*?

Figure 2 constitutes a counter-example. Figure 2(a) plots the singular values of a signal-plus-noise data matrix modeled as $\tilde{X} = 2uv^H + X$, where the noise-only matrix X is a



(a) MCA example.



(b) MCA example.

FIGURE 2. The singular value spectrum of the signal-plus-noise data matrix exhibits two continuous-looking portions that may be associated with “noise” singular values and a single separated singular value that may be interpreted as evidence of a rank-one “signal” matrix buried in the data matrix. Note that in contrast to Figure 1, the principal component is *not* the most informative component.

mixture of two multivariate Gaussians with different variances that produces a spectrum that is supported on two disconnected intervals. The MATLAB code used to generate \tilde{X} is listed below so the reader may reproduce Figure 2:

```
n = 1000; m = n;
Sigma = diag([20*ones(n/10,1);ones(n-n/10,1)],0); % temporal covariance
G = randn(n,m)/sqrt(m)*sqrtm(Sigma);
u = randn(n,1); u = u/norm(u); v = randn(m,1)/sqrt(m);
Xtil = 2*u*v' + G;
```

The presence of a signal matrix is reflected in the single singular value that separates from the continuous looking portions of the spectrum - unlike Figure 1(a), it is in the

middle *i.e.* **not** associated with the principal component that captures the greatest variation. The rule in (2) would return $\hat{r} = 0$ here and we would fail to detect the underlying signal matrix.

Figure 2(b) plots the inner-product $\{|\langle \tilde{u}_i, u \rangle|^2\}_{i=1}^n$, where $\{\tilde{u}_i\}_{i=1}^n$ are the left singular vectors of \tilde{X} . The quantities $\{|\langle \tilde{u}_i, u \rangle|^2\}_{i=1}^n$ (and $\{|\langle \tilde{v}_i, v \rangle|^2\}_{i=1}^m$) are measures of informativeness of the singular vectors of \tilde{X} with respect to the singular vectors of the latent signal matrix. Clearly, the principal left (also, the right - not plotted here) singular vector is **not** the most informative component; the middle component is. Employing the principal component in an estimate of the signal matrix as in (3) would not be as judicious as using the most informative component, which is the *middle component* here.

The preceding examples support our assertion that the principal components are not necessarily the most informative components and that middle components might sometimes be more important. The examples also hint at the role played by the spectrum of the noise-only matrix X in determining the relative informativeness of the components.

An additional remark is in order. The Eckart-Young-Mirsky (EYM) theorem [10, 23] states that for any unitarily invariant norm, the optimal rank \hat{r} approximation to \tilde{X}_n is given by (3). This is a statement about optimal *representation* of the signal-plus-noise matrix. It is **not** a statement about inference on the underlying low-rank signal matrix. Thus there is no contradiction between our results and the content of the EYM theorem.

1.1. Motivation and summary of findings. This work is motivated by the ubiquity of principal component analysis (PCA) in data analysis and signal processing and the associated importance assigned by practitioners to the leading singular values and vectors of the data matrix.

In emerging applications, such as the collaborative learning, graph mining or bioinformatics where the data matrix is large, it is infeasible to compute the entire singular value decomposition. There are, however, efficient techniques for computing the leading singular vectors of a matrix that employ iterative techniques such as the Arnoldi or Lanczos iteration [6] and the family of Krylov subspace methods or using randomized techniques as in [13, 9, 8, 14].

In these ‘big data’ applications, researchers often invoke PCA as justification for the computation of a small number of leading singular vectors of the data matrix. Arguably, what a practitioner who uses these principal components as a starting point in an inferential detection, estimation or classification procedure is really after are the informative components. As we have already seen, the informative components need not be the principal components and may even be the middle components.

In the latter scenario, computation of the leading singular vectors, regardless of computational considerations or choice of algorithm, might lead to faulty inference and lead a non-specialist down a road to a flawed conclusion that they may present as supported by standard PCA derived data analysis. The situation is particularly perilous in biomedical applications involving high-dimensional data sets where one cannot exclude or reason about most informative components by visual inspection.^{2,3}

²<http://www.nytimes.com/2011/07/19/health/19gene.html?pagewanted=all>

³http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=2&hp

A first-principles approach is needed to justify why the principal components might be informative for simple, canonical noise models but also for identifying when middle components might be informative. This paper is a step in that direction. In what follows, we provide a complete picture of how the spectrum of X governs the informativeness of various components of the SVD of a data matrix \tilde{X} modeled as in (1). To summarize our findings:

- The informative components correspond to isolated singular values that separate from the noise (or continuous looking) component of the spectrum,
- Principal components are the most informative components when the noise (or the continuous looking) component of the spectrum is supported on one interval,
- Middle components may be informative when the noise component of the spectrum is supported on multiple intervals,
- Heterogeneities in the noise-only matrix can produce a disconnected noise spectrum,
- It is possible for both principal and middle components to be informative and,
- It is possible for the middle component to be informative even when the principal component is uninformative.

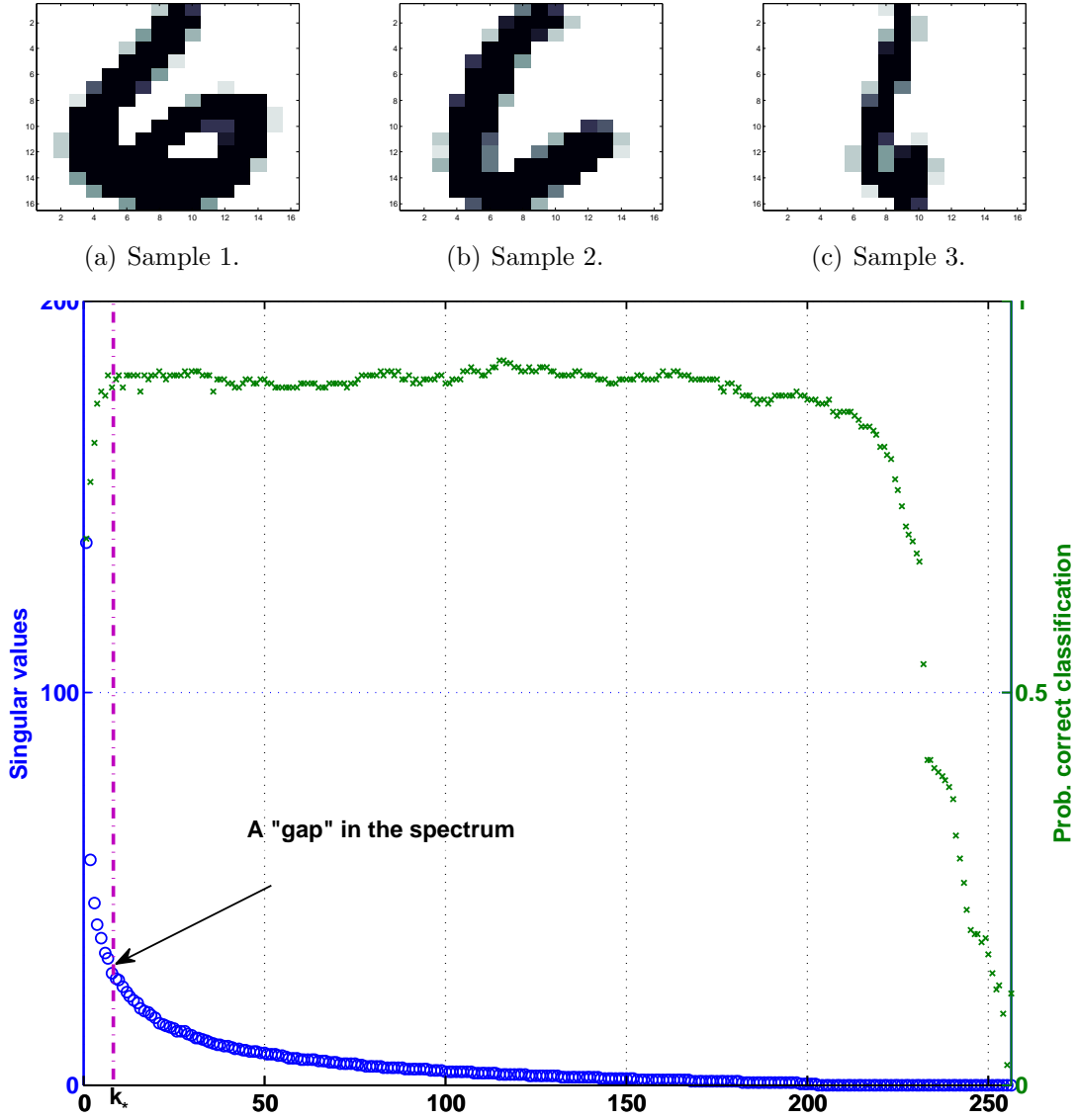
Our findings will allow the practitioner to better justify, by employing reasoning based on the *entire* spectrum of X , when the use of principal components is warranted (as it is for the example in Figure 3) and when the middle components might be more informative as in Figure 2. The next step in this line of inquiry, that is beyond the scope of this paper, is the development of efficient computational methods for large data sets that can detect and extract informative middle components.

We conclude by submitting Figure 4 as evidence that our findings describe phenomena that might already be present in real-world data sets ⁴ that might previously have been interpreted differently. Here we have a 438×1200 data matrix whose columns contains measurements made at a receiver sensor array and some of the past transmitted data symbols. The measurements were made over a time period where there were significant fluctuations in the noise levels. The fluctuations in the channel transfer function constitute the low-rank “signal” here.

The plot of the singular values in Figure 4 contains clusters of principal and middle eigenvalues that separate from the continuous looking portion of the spectrum. Our findings suggest that these are informative principal and middle components. We hope that this work contributes to an increased understanding of the role played by the noise eigen-spectrum in shaping the informativeness of various SVD components and a recognition that there is much left to understand in terms of low-rank signal extraction from noisy data matrices.

We begin our exposition in Section 2 by examining how the spectrum of \tilde{X} is related to the spectrum of X . We utilize the findings in Section 3 to analyze a setting where the principal components are informative. In Section 4 we describe a scenario when middle components can be informative while Section 5 contains the main results which formalize the arguments presented in Sections 3 and 4. We conclude in Section 6 with a discussion of which noise models can produce informative middle and principal components.

⁴ We thank Dr. James Preisig of the Woods Hole Oceanographic Institution for this dataset.



(d) The singular value spectrum of the training data matrix for the digit “6” is shown.

FIGURE 3. (a) - (c) Three samples representing the digit “6” from the USPS handwritten digits database. Each s -pixel-by- s -pixel training image is converted into a $n = s^2 \times 1$ column vector whose elements represent grayscale values. The training data matrix is formed by stacking the column vectors corresponding to every image in the labeled training data set alongside each other. (d) displays on the singular values of the data matrix on the left axis. As in Figure 1(a), the singular value spectrum exhibits a continuous-looking portion (that may be interpreted as “noise”) and a separated portion (that may be interpreted as low-rank “signal”). (right axis) A plot of a probability of correct classification versus r plot where r is the number of left singular vectors of the training set used for classification. *Note that choosing r based on the singular value “gap separation” heuristic yields near-optimal performance.*

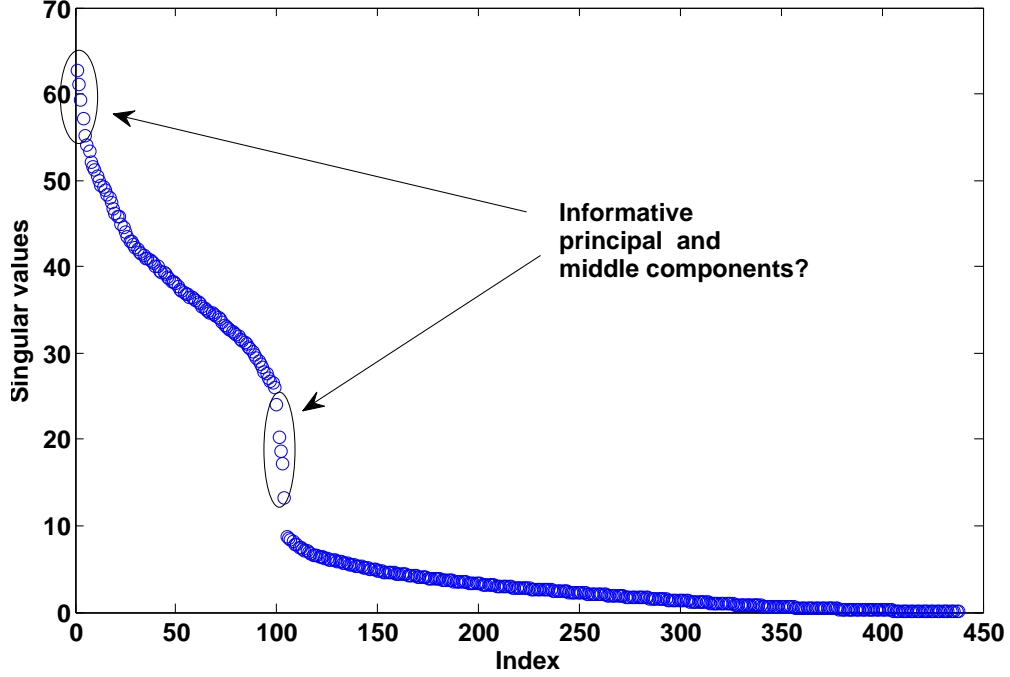


FIGURE 4. A real-world data set with possibly informative principal and middle components.

2. THE EIGENVALUES AND EIGENVECTORS OF \tilde{X}

For expositional simplicity, let us consider the model in (1) with $r = 1$ and symmetric \tilde{X} , so that

$$\tilde{X} = S + X,$$

where $S = \theta uu^*$, for some arbitrary, non-random, unit norm column vector u . We begin our investigation by examining how the eigenvalues and eigenvectors of \tilde{X} are related to the eigenvalues and eigenvectors of the low-rank signal matrix S .

Let $X = Q\Lambda Q^*$ be the eigen-decomposition of the noise-only random matrix X (we have suppressed the subscript in X_n for notational brevity), where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ and Q are the eigenvectors of X . We assume that the noise-only random matrix X is invariant, in distribution, under orthogonal (or unitary) conjugation. This implies that the eigenvectors of X are Haar-distributed and independent of its eigenvalues [15, Th. 4.3.5]. We will utilize this fact shortly.

2.1. Eigenvalues of \tilde{X} . The eigenvalues of $X + S$ are the solutions of the equation

$$\det(zI - (X + S)) = 0.$$

Equivalently, for z such that $zI - X$ is invertible, we have

$$zI - (X + S) = (zI - X) \cdot (I - (zI - X)^{-1}S),$$

so that

$$\det(zI - (X + S)) = \det(zI - X) \cdot \det(I - (zI - X)^{-1}S).$$

Consequently, a simple argument reveals that the z is an eigenvalue of $X + S$ and not an eigenvalue of X if and only if 1 is an eigenvalue of the matrix $(zI - X)^{-1}S$. But $(zI - X)^{-1}S = (zI - X)^{-1}\theta uu^*$ has rank one, so its only non-zero eigenvalue will equal its trace, which in turn is equal to $\theta u^*(zI - X)^{-1}u = \theta u^*Q(zI - \Lambda)^{-1}Q^*u$.

Let $v = Q^*u$. Then, z is an eigenvalue of \tilde{X} and not an eigenvalue of X if and only if

$$\sum_{i=1}^n \frac{|v_i|^2}{z - \lambda_i} = \frac{1}{\theta}. \quad (4)$$

Let μ_n be the “weighted” spectral measure of X , defined by

$$\mu_n = \sum_{i=1}^n |v_i|^2 \delta_{\lambda_i} \quad (\text{the } v_i\text{'s are the coordinates of } v = Q^*u). \quad (5)$$

Then any z outside the spectrum of X is an eigenvalue of \tilde{X} if and only if

$$\sum_{i=1}^n \frac{|v_i|^2}{z - \lambda_i} =: G_{\mu_n}(z) = \frac{1}{\theta}, \quad (6)$$

where $G_{\mu}(z)$ is the Cauchy transform of μ defined as

$$G_{\mu}(z) = \int \frac{1}{z - x} d\mu(x). \quad (7)$$

Equation (6) describes the *exact* relationship between the eigenvalues of \tilde{X} and the eigenvalues of X and the dependence on the coordinates of the vector v (via the measure μ_n), which we will use shortly.

2.2. Eigenvectors of \tilde{X} . Let \tilde{u} be a unit eigenvector of $X + S$ associated with the eigenvalue z that satisfies (6). From the relationship $(X + S)\tilde{u} = z\tilde{u}$, we deduce that, for $S = \theta uu^*$,

$$(zI - X)\tilde{u} = S\tilde{u} = \theta uu^*\tilde{u} = (\theta u^*\tilde{u})u \quad (\text{because } u^*\tilde{u} \text{ is a scalar}),$$

implying that \tilde{u} is proportional to $(zI - X)^{-1}u$.

Since \tilde{u} has unit-norm,

$$\tilde{u} = \frac{(zI - X)^{-1}u}{\sqrt{u^*(zI - X)^{-2}u}} \quad (8)$$

and

$$|\langle u, \tilde{u} \rangle|^2 = |u^*\tilde{u}|^2 = \frac{(u^*Q(zI - \Lambda)^{-1}Q)^2}{u^*Q(zI - \Lambda)^{-2}Q^*u} = \frac{G_{\mu_n}(z)^2}{\int \frac{d\mu_n(x)}{(z-x)^2}} = \frac{1}{\theta^2 \int \frac{d\mu_n(x)}{(z-x)^2}}. \quad (9)$$

Notice that

$$\int \frac{d\mu_n(x)}{(z-x)^2} = -G'_{\mu_n}(z) \quad (10)$$

so that we have

$$|\langle u, \tilde{u} \rangle|^2 = -\frac{1}{\theta^2 G'_{\mu_n}(z)} \quad (11)$$

Equation (8) describes the relationship between the eigenvectors of \tilde{X} and the eigenvalues of X and the dependence on the coordinates of the vector v (via the measure μ_n), which we will return to shortly.

3. WHEN PRINCIPAL COMPONENTS ARE THE MOST INFORMATIVE COMPONENTS

We begin our investigation by considering a setting where the informative components do indeed correspond to the principal components. The picture we have developed so far is that the eigenvalues z_i and the associated eigenvectors \tilde{u}_i of the signal-plus-noise data matrix \tilde{X} modeled as $\tilde{X} = X + \theta uu^*$ satisfy the equations

$$G_{\mu_n}(z_i) = \frac{1}{\theta}, \quad (12a)$$

$$|\langle \tilde{u}_i, u \rangle|^2 = -\frac{1}{\theta^2} \cdot \frac{1}{G'_{\mu_n}(z_i)}, \quad (12b)$$

where

$$G_{\mu_n}(z) = \sum_{i=1}^n \frac{|v_i|^2}{z - \lambda_i}.$$

The expressions in (12) provide insight on how the eigenvalues of \tilde{X} are related to the eigenvalues of X .

Figure 5 considers the $n = 5$ setting and shows how the expressions in (12) provide insight on the informativeness of the eigenvalues and eigenvectors of \tilde{X} .

By (12a), the eigenvalues of \tilde{X} correspond to the values of z where the horizontal line $1/\theta$ in Figure 5 intersects the curve $G_{\mu_n}(z)$. Since $G_{\mu_n}(z)$ has poles at the eigenvalues of X , all but the largest eigenvalue of \tilde{X} interlace the eigenvalues of X . Consequently, $\lambda_5 \leq \tilde{\lambda}_5 \leq \lambda_4$ and so on; there is no eigenvalue to the right of λ_1 and hence $\tilde{\lambda}_1$ can be displaced by a greater amount, subject to $\tilde{\lambda}_1 - \lambda_1 \leq \theta$.

Equation (12b) reveals that the informativeness of an eigenvector, denoted by $\text{Inf}_i := |\langle \tilde{u}_i, u \rangle|^2$, is inversely proportional to the negative slope of the function $G_{\mu_n}(z)$ evaluated at the eigenvalue $z_i = \tilde{\lambda}_i$ of \tilde{X} associated with the eigenvector \tilde{u}_i .

3.1. Asymptotic analysis: Eigenvalues. We now place ourselves in the high-dimensional setting. Let us assume that as $n \rightarrow \infty$,

$$\mu_{X_n} = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i} \xrightarrow{\text{a.s.}} \mu_X,$$

where μ_X is a non-random probability and $\xrightarrow{\text{a.s.}}$ denotes almost sure convergence⁵. Assume that the largest and smallest eigenvalues of X_n converge to b and a , respectively and that $d\mu_X(z) > 0$ for all $z \in (a, b)$ so that the measure is supported on one connected interval. When X is a sample covariance matrix formed from a matrix with i.i.d. Gaussian variables, the eigenvalues will satisfy this condition [29].

⁵The argument holds for other modes of convergence as well so we shall not explicitly specify the mode of convergence in the expository sections that follow.

The assumed convergence of the eigenvalues to a smooth limiting measure implies that as $n \rightarrow \infty$, if there were no signal, the eigenvalues would have a continuous looking spectrum as the spacing between successive eigenvalues goes to zero. By the same reasoning, when there is a signal, the picture developed in Figure 5 says that all but the leading eigenvalue of \tilde{X} will be displaced insignificantly. Thus the $n - 1$ eigenvalues will retain their continuous looking nature and will be tightly packed together.

As $n \rightarrow \infty$, only the largest eigenvalue will exhibit a significant $O(1)$ deviation relative to the corresponding eigenvalue in the noise-only setting (*i.e.*, when $S = 0$). Since the second largest eigenvalue is also displaced insignificantly by a vanishing (with n) amount, this manifests as an $O(1)$ gap in the spectrum as in Figure 1(a) and the use of the (principal) gap heuristic in (2) for signal detection is justified.

We now investigate the fundamental limit of gap heuristic based signal detection. We first note that the vector $v = Q^*u$ is uniformly distributed on the unit hypersphere, and so, in the high-dimensional setting, $|v_i|^2 \approx 1/n$ (with high probability) so that

$$\sum_{i=1}^n |v_i|^2 \delta_{\lambda_i} =: \mu_n \approx \mu_X := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}.$$

A consequence of $\mu_n \rightarrow \mu_X$ is that $G_{\mu_n}(z) \rightarrow G_{\mu_X}(z)$. Inverting equation (6) after substituting these approximations yields the location of the largest eigenvalue, in the $n \rightarrow \infty$ limit to be $G_{\mu_X}^{-1}(1/\theta)$.

Recall that we had assumed that the limiting probability measure of the noise-only random matrix μ_X is compactly supported on a single, connected interval $[a, b]$. Consequently, the Cauchy transform G_{μ_X} given by (7) is well-defined for z *outside* $[a, b]$ and can tend to a limit $G_{\mu_X}(b^+)$ which may be bounded, *i.e.* have $G_{\mu_X}(b^+) < +\infty$.

So long as $1/\theta < G_{\mu_X}(b^+)$, as in Figure 6-(a), we obtain $\lambda_1(\tilde{X}) \approx G_{\mu_X}^{-1}(1/\theta) > b$. This results in an $O(1)$ gap between the largest eigenvalue and the edge of the spectrum and the gap heuristic will work. However, when $1/\theta \geq G_{\mu_X}(b^+)$, as in Figure 6-(b), $\lambda_1(\tilde{X}) \rightarrow \lambda_1(X) = b$ and the gap heuristic will fail. To summarize:

Principal gap based signal detection will asymptotically succeed iff $\theta > 1/G_{\mu_X}(b^+)$.

3.2. Asymptotic analysis: Eigenvectors. Recall our argument that since Q is isotropically random, the vector $v = Q^*u$ is uniformly distributed on the unit hypersphere and $|v_i|^2 \approx 1/n$ (with high probability) in the high-dimensional setting. Consequently, we have that

$$G'_{\mu_n}(z) = -\frac{|v_i|^2}{(z - \lambda_i)^2} \approx -\frac{1}{n} \sum_{i=1}^n \frac{1}{(z - \lambda_i)^2}.$$

Since all the n eigenvalues of the noise-only matrix are concentrated on the connected interval $[a, b]$, the average spacing between the eigenvalues of X is $O(1/n)$. Since the eigenvalues of \tilde{X} interlace the eigenvalues of X , $\tilde{\lambda}_i - \lambda_i = O(1/n)$ for all but the largest eigenvalue. Hence $G'_{\mu_n}(z) = O(n)$ so that $\{\text{Inf}_i\}_{i=2}^n = -1/G'_{\mu_n}(z)|_{z=\tilde{\lambda}_i} = O(1/n)$.

However, $\tilde{\lambda}_1 - \lambda_1 = O(1)$, so that $G'_{\mu_n}(\tilde{\lambda}_1) = O(1)$ and $\ln f_1 = O(1)$ implying that the principal eigenvector is maximally informative with a non-vanishing (with n) informativeness and the use of (3) in the estimation of S is justified. We now investigate the fundamental limit of principal eigenvector based signal estimation.

In the asymptotic setting when $\mu_n \rightarrow \mu_X$ and $z = \tilde{\lambda}_1 \rightarrow \rho$ we have that

$$\int \frac{d\mu_n(t)}{(z-t)^2} \rightarrow \int \frac{d\mu_X(t)}{(\rho-t)^2} = -G'_{\mu_X}(\rho),$$

so that when $1/\theta < G_{\mu_X}(b^+)$, which implies that $\rho > b$, we have

$$|\langle \tilde{u}_1, u \rangle|^2 \xrightarrow{\text{a.s.}} \frac{1}{\theta^2 \int \frac{d\mu_X(t)}{(\rho-t)^2}} = \frac{-1}{\theta^2 G'_{\mu_X}(\rho)} > 0,$$

whereas when $1/\theta \geq G_{\mu_X}(b^+)$ and if μ_X is such that G_{μ_X} has infinite derivative at $\rho = b$, we have

$$|\langle \tilde{u}_1, u \rangle| \xrightarrow{\text{a.s.}} 0.$$

Hence when $\theta \leq 1/G_{\mu_X}(b^+)$ and if $G'_{\mu_X}(b^+) = \infty$, then the all components have vanishing (with n) informativeness. To summarize, when $\theta > 1/G_{\mu_X}(b^+)$:

Principal components are the most informative components when the noise eigen-spectrum is contained on a single, connected interval .

The eigen-spectrum of a Wishart distributed sample covariance matrix with identity covariance satisfies this condition. It is thus a happy coincidence that principal components are the most informative components for the simplest noise matrix model. We now consider the setting where the noise eigen-spectrum is (asymptotically) supported on multiple disconnected intervals.

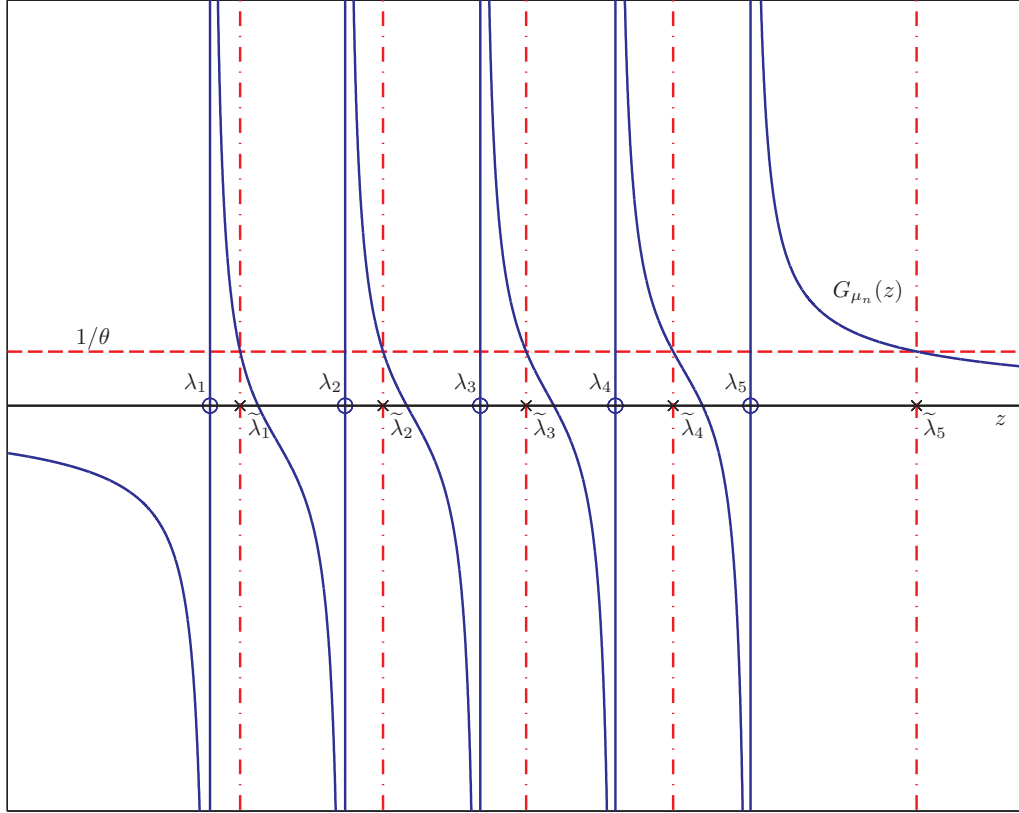
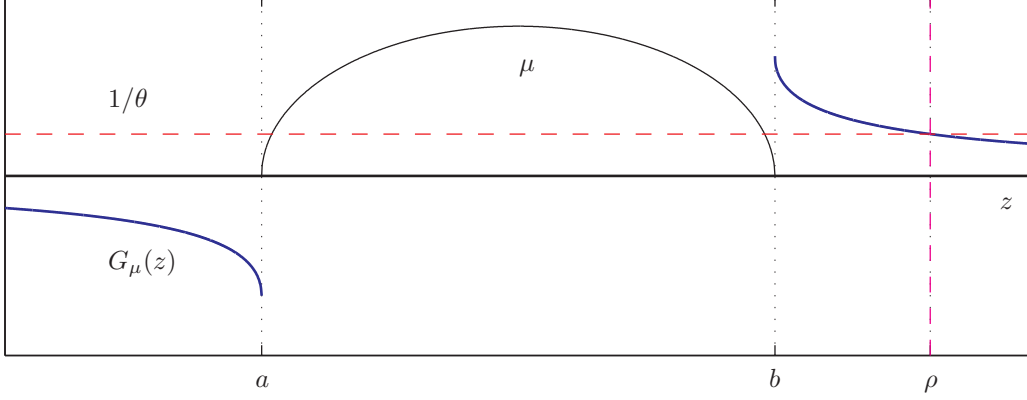
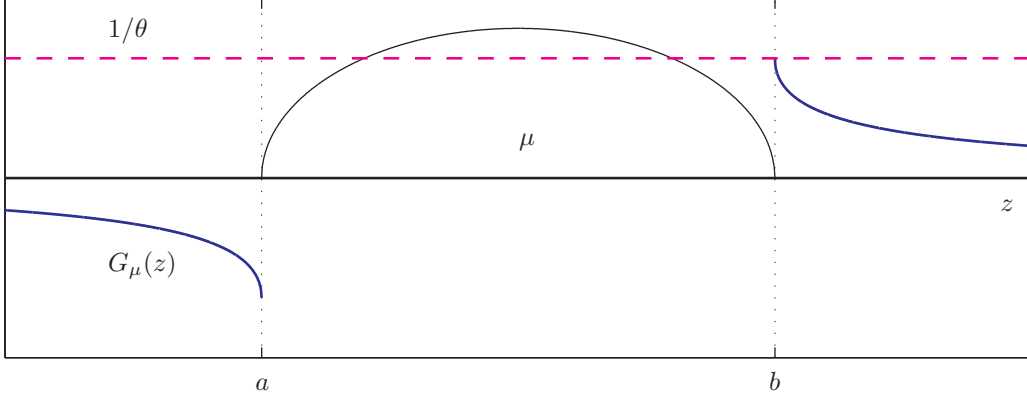


FIGURE 5. The relationship in (12a) between the eigenvalues of $\tilde{X} = \theta uu^* + X$ and the eigenvalues of X is depicted here. Notice the interlacing of the bulk eigenvalues and the emergence of the principal eigen-gap.



(a) When $1/\theta < G_\mu(b)$, $\lambda_1(\tilde{X}) \rightarrow \rho = G_\mu^{-1}(\frac{1}{\theta})$ and there is a principal eigen-gap.



(b) When $1/\theta = G_\mu(b)$, $\lambda_1(\tilde{X}) \rightarrow b$ and there is no principal eigen-gap.

FIGURE 6. The evolution of the informativeness of the principal eigen-gap for different values of θ . In (a), where θ is large, the principal eigen-gap is informative; (b) the principal eigen-gap vanishes when $1/\theta = G_\mu(b^+)$ and the signal is undetectable using principal eigen-gap based methods.

4. WHEN MIDDLE COMPONENTS ARE INFORMATIVE

4.1. Asymptotic analysis: eigenvalues. Consider a setting where the eigenvalues of the noise-only matrix X are supported on multiple intervals as in Figure 2. This corresponds to letting the μ_X obtained in the $n \rightarrow \infty$ limit of μ_{X_n} being supported on $\ell = 2$ intervals. Thus we may model μ_X as

$$\mu_X = p_1\mu_1 + p_2\mu_2,$$

where $p_1 + p_2 = 1$. Here $p := p_1 \in (0, 1)$ and the measures μ_1 and μ_2 are non-random probability measures supported on $[a_1, b_1]$ and $[a_2, b_2]$, respectively with $d\mu_i(z) > 0$ for $z \in (a_i, b_i)$ for $i = 1, 2$. We suppose that $a_2 < b_2 < a_1 < b_1$, as depicted in Figure 8-(a). For $k = 0, 1, 2$, define $c_j = \sum_{i=0}^j p_i$ with $p_0 := 0$, $c_0 := 0$ and $c_2 := 1$. We assume that for $j = 1, 2$, $\lambda_{nc_{j-1}+1} \xrightarrow{\text{a.s.}} b_j$ and that $\lambda_{nc_j} \xrightarrow{\text{a.s.}} a_j$. For expositional simplicity we assume that nc_j is an integer. When X is a sample covariance matrix formed from a matrix with Gaussian entries having a covariance matrix with an adequately-separated covariance eigen-spectrum then the sample eigenvalues will satisfy this condition. Section 6 contains additional examples and elaborates on when the covariance eigen-spectrum is separated enough.

The assumed convergence of the eigenvalues to a smooth limiting measure implies that as $n \rightarrow \infty$, if there were no signal, the eigenvalues would have a continuous looking spectrum as the spacing between successive eigenvalues goes to zero.

By the same reasoning, when there is a signal, the picture developed in Figure 5 when adapted as in Figure 7 for the disjoint interval setting (here $\ell = 2$) reveals (via (12)) that the leading eigenvalue $\tilde{\lambda}_1$ **and** an additional, middle, eigenvalue $\tilde{\lambda}_{nc_1+1}$ will exhibit a significant $O(1)$ deviation relative to the corresponding eigenvalue in the noise only setting. The middle eigenvalue emerges from the bulk spectrum because $\lambda_{nc_1} - \lambda_{nc_1+1} = O(1)$ as $n \rightarrow \infty$.

The remaining $n - 2$ eigenvalues will be displaced insignificantly and will remain tightly packed together, thereby retaining their continuous looking appearance. Consequently, there will be two $O(1)$ eigen-gaps in the spectrum betraying the presence of a low-rank signal. Thus, here too, the use of the gap heuristic is justified.

The emergence of an informative middle eigenvalue in this setting due to the presence of a large gap in the noise eigen-spectrum may be viewed as a form of aliasing.

We now investigate the fundamental limit of gap heuristic based signal detection so we might understand when not accounting for the middle eigen-gap might lead to suboptimal detection performance.

As before, we note that the vector $v = Q^*u$ is uniformly distributed on the unit hypersphere, and so, in the high-dimensional setting, $|v_i|^2 \approx 1/n$ (with high probability) so that

$$\sum_{i=1}^n |v_i|^2 \delta_{\lambda_i} =: \mu_n \approx \mu_X := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i}.$$

A consequence of $\mu_n \rightarrow \mu_X$ is that $G_{\mu_n}(z) \rightarrow G_{\mu_X}(z)$. Inverting equation (6) after substituting these approximations yields the location of the largest eigenvalue, in the $n \rightarrow \infty$ limit to be $G_{\mu_X}^{-1}(1/\theta)$. This results in multiple (i.e. principal and middle) eigenvalues

that separate from the bulk spectrum precisely when the functional inverse is multi-valued (for a domain outside the region of support).

Recall our assumption that the limiting probability measure of the noise-only random matrix μ_X is compactly supported on $\ell = 2$ disjoint intervals $\{[a_i, b_i]\}_{i=1}^2$. Consequently, the Cauchy transform G_{μ_X} given by (7) is well-defined for z outside $\cup_{i=1}^2 [a_i, b_i]$ and is strictly decreasing with increasing z on open intervals $(\cup_{i=1}^2 [a_i, b_i])^c$ outside the support of μ_X , as depicted in Figure 8.

Thus so long as $1/\theta < G_{\mu_X}(b_1^+)$, $\lambda_1 \rightarrow \rho_1 > b_1$ and an $O(1)$ principal eigen-gap will manifest. Conversely, if $1/\theta \geq G_{\mu_X}(b_1^+)$, as in Figure 8-(b), $\lambda_1 \rightarrow b_1$ and there will be no principal eigen-gap.

Similarly, if $1/\theta < G_{\mu_X}(b_2^+)$ and $1/\theta > G_{\mu_X}(a_1^-)$, $\lambda_{nc_1+1} = \lambda_{np+1} \rightarrow \rho_2 > b_2$ and an $O(1)$ middle eigen-gap will manifest. Conversely, as shown in Figure 8-(c), if $1/\theta > G_{\mu_X}(b_2^+)$ then $\lambda_{nc_1+1} = \lambda_{np+1} \rightarrow b_2$ and there will be no $O(1)$ middle eigen-gap. However when $1/\theta < G_{\mu_X}(a_1^-)$, then $\lambda_{nc_1+1} = \lambda_{np+1} \rightarrow a_1$ and technically speaking there is an $O(1)$ middle eigen-gap except that this gap is indistinguishable from the gap in the spectrum that appears even when there is no signal.

Thus principal eigen-gap based signal detection for weak signals (or small θ) fails whenever $\theta < 1/G_{\mu_X}(b_1^+)$ while middle eigen-gap detection fails whenever $\theta < 1/G_{\mu_X}(b_2^+)$. If $G_{\mu_X}(b_2^+) > G_{\mu_X}(b_1^+)$, as depicted in Figure 8, then a weak signal that is undetectable using the principal eigen-gap heuristic would have remained detectable if the middle eigen-gap were considered. This is why the middle eigen-gap in Figure 2 was informative while the principal eigen-gap was not. In such settings, detection using **only** the principal eigen-gap detection is suboptimal.

4.2. Asymptotic analysis: eigenvectors. Equation (12b) reveals that the informativeness of an eigenvector \tilde{u}_i , relative to the signal eigenvector u is given by the expression

$$|\langle \tilde{u}_i, u \rangle|^2 = -\frac{1}{\theta^2} \cdot \frac{1}{G'_{\mu_n}(z_i)},$$

where

$$G'_{\mu_n}(z) = -\frac{|v_i|^2}{(z - \lambda_i)^2} \approx -\frac{1}{n} \sum_{i=1}^n \frac{1}{(z - \lambda_i)^2}.$$

The eigenvalues of the noise-only matrix are concentrated on the disjoint intervals $[a_1, b_1]$ and $[a_2, b_2]$. Thus, the average spacing between the successive eigenvalues of X within each interval is $O(1/n)$. Since the eigenvalues of \tilde{X} interlace the eigenvalues of X , $\tilde{\lambda}_i - \lambda_i = O(1/n)$ for all but the largest eigenvalue and the middle eigenvalue as in Figure 8-(a). Hence $G'_{\mu_n}(z) = O(n)$ so that $\{\text{Inf}_i\}_{i=2}^n = 1/G'_{\mu_n}(z)|_{z=\tilde{\lambda}_i} = O(1/n)$.

As before, we note that $\tilde{\lambda}_1 - \lambda_1 = O(1)$ so that $G'_{\mu_n}(\tilde{\lambda}_1) = O(1)$ and $\text{Inf}_1 = O(1)$ implying that the principal eigenvector is informative with a non-vanishing (with n) informativeness and the use of (3) in the estimation of S is justified. However, what emerges from the picture in Figure 7 is that since $\lambda_{np} - \lambda_{np+1} = O(1)$ we have that $\tilde{\lambda}_{np+1} - \lambda_{np+1} = O(1)$ and by the same argument, $\text{Inf}_{np+1} = O(1)$ as well. Thus the middle eigenvector associated with the middle eigenvalue that exhibits an eigen-gap is also informative. Employing it in the estimation of S in (3) would improve estimation performance.

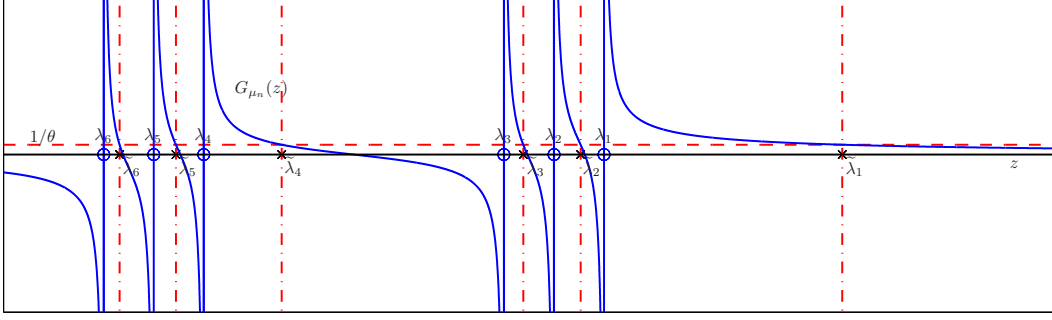


FIGURE 7. The relationship in (12a) between the eigenvalues of $\tilde{X} = \theta uu^* + X$ and those of X is depicted here when the eigenvalues of X are supported on two $O(1)$ separated intervals. Notice the interlacing of the bulk eigenvalues and the emergence of the principal and middle eigen-gap. Contrast this to Figure 5 when only the principal eigen-gap emerges.

Extending this argument further, in the $n \rightarrow \infty$ limit, when $\mu_n \rightarrow \mu_X$ suppose μ_X is such that $G'_{\mu_X}(z) = -\infty$ for $z = b_1, b_2$. Then we have that whenever $1/\theta < G_{\mu_X}(b_1^+)$,

$$|\langle \tilde{u}_1, u \rangle|^2 \xrightarrow{\text{a.s.}} \frac{1}{\theta^2 \int \frac{d\mu_X(t)}{(\rho_1 - t)^2}} = \frac{-1}{\theta^2 G'_{\mu_X}(\rho_1)} > 0,$$

but if $1/\theta \geq G_{\mu_X}(b_1^+)$ as in Figure 8-(b),(c), then $|\langle \tilde{u}_1, u \rangle| \xrightarrow{\text{a.s.}} 0$, and the principal component becomes uninformative. Employing the same argument for the middle eigenvector reveals that so long as $G_{\mu_X}(a_1^-) \leq 1/\theta < G_{\mu_X}(b_2^+)$ then $\lambda_{np+1} \xrightarrow{\text{a.s.}} \rho_2$ and the corresponding eigenvector is informative *i.e.*,

$$|\langle \tilde{u}_{np+1}, u \rangle|^2 \xrightarrow{\text{a.s.}} \frac{-1}{\theta^2 G'_{\mu_X}(\rho_2)} > 0.$$

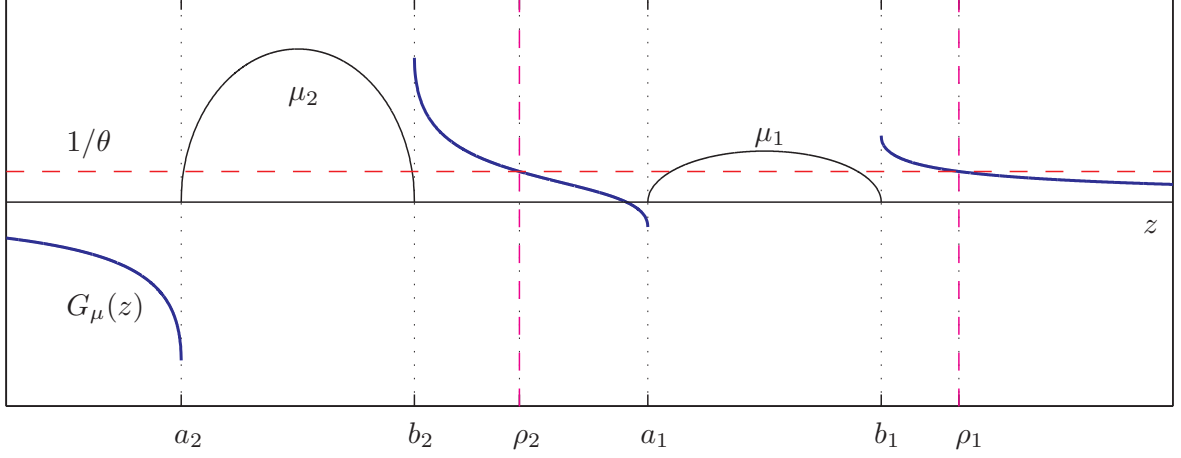
When $1/\theta \geq G_{\mu_X}(b_2^+)$ as in Figure 8-(c), then

$$|\langle \tilde{u}_1, u \rangle| \xrightarrow{\text{a.s.}} 0,$$

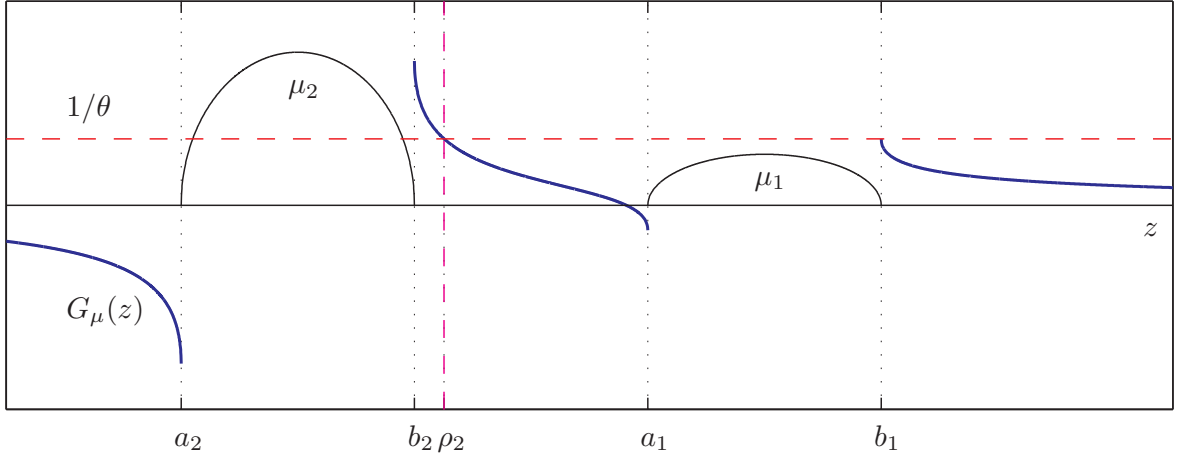
and the middle component becomes uninformative. Evidently, if $1/G_{\mu_X}(b_2^+) > 1/G_{\mu_X}(b_1^+)$ as in Figure 8 then the middle eigenvector will stay informative for a regime of small θ where the principal eigenvector is uninformative. More generally, if both the principal and the middle eigenvectors are informative then principal eigenvector will be more informative if $-1/G'_{\mu_X}(\rho_1) > -1/G'_{\mu_X}(\rho_2)$ and vice versa. This is determined by the structure of the noise spectrum. To summarize:

- Principal gap based signal detection will asymptotically succeed iff $\theta > 1/G_{\mu_X}(b_1^+)$,
- Middle gap based detection will asymptotically succeed despite principal gap based detection failing whenever $G_{\mu_X}(b_2^+) > G_{\mu_X}(b_1^+)$.
- The eigenvectors associated with principal or middle eigenvalues that exhibit an eigen-gap will be informative
- The eigenvectors will be uninformative when the eigen-gap vanishes.

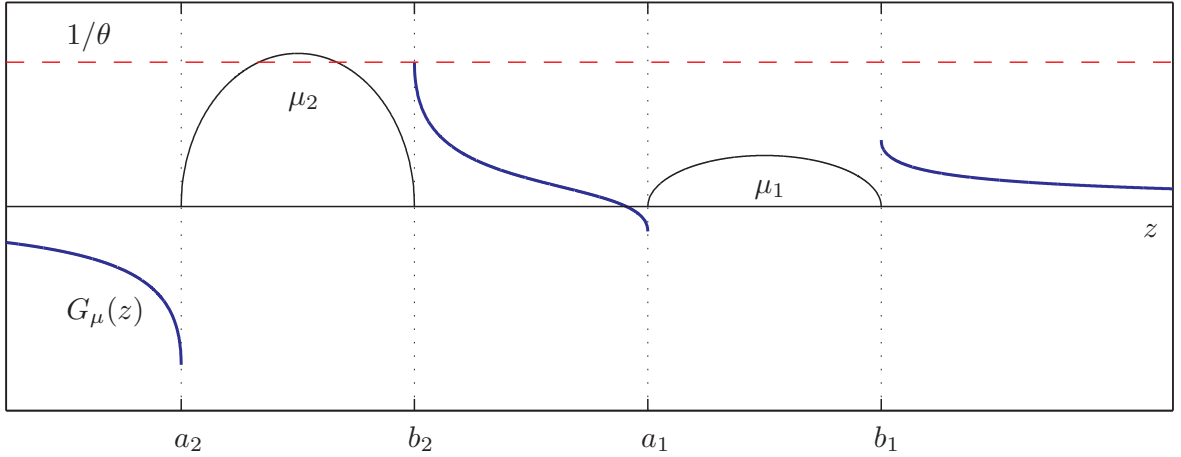
The emergence of informative middle eigenvalues and eigenvector whenever there is a gap in the noise eigen-spectrum may be viewed as a form of signal (subspace) aliasing.



(a) When $1/\theta < G_{\mu_X}(b_1^+)$ and $1/\theta < G_{\mu_X}(b_2^+)$, $\tilde{\lambda}_1(\tilde{X}) \rightarrow \rho_1 > b_1$ and $\tilde{\lambda}_{np+1} \rightarrow \rho_2 > b_2$.



(b) When $1/\theta = G_{\mu_X}(b_1^+)$, $\tilde{\lambda}_1 \rightarrow b_1$. Since $1/\theta < G_{\mu_X}(b_2^+)$, $\tilde{\lambda}_{np+1} \rightarrow \rho_2 > b_2$.



(c) When $1/\theta > G_{\mu_X}(b_1^+)$, $\tilde{\lambda}_1 \rightarrow b_1$. Since $1/\theta = G_{\mu_X}(b_2^+)$, $\tilde{\lambda}_{np+1} \rightarrow b_2$.

FIGURE 8. The evolution of the informativeness of the principal and the middle eigen-gaps for different values of θ . In (a), where θ is large, both the principal and middle eigen-gaps are informative; (b) the principal eigen-gap vanishes but the middle eigen-gap persists; (c) both the principal and middle eigen-gaps vanishes and the signal is undetectable using eigen-gap based methods. The important point to note here is that the middle eigen-gap reveals the presence of a signal even when the principal eigen-gap does not.

5. MAIN RESULTS

5.1. Eigenvalues and Eigenvectors. Let X_n be an $n \times n$ symmetric (or Hermitian) random matrix whose ordered eigenvalues we denote by $\lambda_1(X_n) \geq \dots \geq \lambda_n(X_n)$. Let μ_{X_n} be the empirical eigenvalue distribution, *i.e.*, the probability measure defined as

$$\mu_{X_n} = \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(X_n)}.$$

Assume that the probability measure μ_{X_n} converges almost surely weakly, as $n \rightarrow \infty$, to a non-random compactly supported probability measure μ_X that is supported on ℓ disjoint intervals so that

$$\mu_X = \sum_{j=1}^{\ell} p_j \mu_j,$$

where for $j = 1, \dots, \ell$, the measures $\mu_j(x)$ is a non-random probability measure are supported on $[a_j, b_j]$ with $d\mu_j(z) > 0$ for $z \in (a_j, b_j)$ and $a_\ell < b_\ell < a_{\ell-1} < b_{\ell-1} < \dots < a_1 < b_1$. Define $c_k = \sum_{i=0}^k p_i$ with $p_0 := 0$, $c_0 := 0$ and $c_\ell := 1$. We assume that for $j = 1, \dots, \ell$, $\lambda_{[nc_{j-1}]+1} \xrightarrow{\text{a.s.}} b_j$ and that $\lambda_{[nc_j]} \xrightarrow{\text{a.s.}} a_j$, where $[nc_{j-1}]$ denotes the smallest integer greater than or equal to nc_{j-1} .

For a given $r \geq 1$, let $\theta_1 \geq \dots \geq \theta_r$ be deterministic non-zero real numbers, chosen independently of n . For every n , let P_n be an $n \times n$ symmetric (or Hermitian) random matrix having rank r with its r non-zero eigenvalues equal to $\theta_1, \dots, \theta_r$.

Recall that a symmetric (or Hermitian) random matrix is said to be *orthogonally invariant* (or *unitarily invariant*) if its distribution is invariant under the action of the orthogonal (or unitary) group under conjugation.

We suppose that X_n and P_n are independent and that X_n , the noise-only, matrix is unitarily invariant while the low-rank signal matrix P_n is non-random.

5.1.1. Notation. Throughout this paper, for f a function and $c \in \mathbb{R}$, we set

$$f(c^+) := \lim_{z \downarrow c} f(z); \quad f(c^-) := \lim_{z \uparrow c} f(z),$$

we also let $\xrightarrow{\text{a.s.}}$ denote almost sure convergence. The ordered eigenvalues of an $n \times n$ Hermitian matrix M will be denoted by $\lambda_1(M) \geq \dots \geq \lambda_n(M)$. Lastly, for a subspace F of a Euclidian space E and a vector $x \in E$, we denote the norm of the orthogonal projection of x onto F by $\langle x, F \rangle$.

Consider the rank r additive perturbation of the random matrix X_n given by

$$\tilde{X} = X_n + P_n.$$

For this model, we establish the following results.

Theorem 5.1 (Eigen-gap phase transition). *The eigenvalues of \tilde{X} exhibit the following behavior as $n \rightarrow \infty$. We have that for each $1 \leq i \leq r$ and $1 \leq j \leq \ell$,*

$$\lambda_{[nc_{j-1}] + i}(\tilde{X}) \xrightarrow{a.s.} \begin{cases} G_{\mu_X, (b_j, a_{j-1})}^{-1}(1/\theta_i) & \text{if } 1/G_{\mu_X}(b_j^+) < \theta_i < 1/G_{\mu_X}(a_{j-1}^-), \\ b_j & \text{if } \theta_i < 1/G_{\mu_X}(b_j^+), \\ a_{j-1} & \text{if } \theta_i > 1/G_{\mu_X}(a_{j-1}^-) \end{cases}$$

Here,

$$G_{\mu_X}(z) = \int \frac{1}{z-t} d\mu_X(t) \quad \text{for } z \notin \text{supp } \mu_X,$$

is the Cauchy transform of μ_X , $G_{\mu_X, (b_j, a_{j-1})}^{-1}(\cdot)$ is its functional inverse for $G_{\mu_X}(z)$ for $z \in (b_j, a_{j-1})$ and $a_0 := +\infty$.

Proof. The result is obtained by following the approach taken in [4, pp. 511-514] for proving Theorem 2.1. The key difference is that we are explicitly considering measures μ_X supported on multiple (disconnected) intervals so that the Cauchy transform of μ_X can have multiple inverses as in Figure 8. For those values of θ such that $G_{\mu_X}^{-1}(1/\theta)$ is multi-valued, as many eigenvalues of \tilde{X} as there are values of z such that $z = G_{\mu_X}^{-1}(1/\theta)$ will exhibit the eigen-gaps identified. \square

Theorem 5.2 (Informativeness of the eigenvectors). *Assume throughout that $\theta > 0$ and let $G_{\mu_X}(a_0^-) = +\infty$. Consider $i_0 \in \{1, \dots, r\}$ such that $1/\theta_{i_0} \in \bigcup_{j=1}^l (G_{\mu_X}(a_{j-1}^-), G_{\mu_X}(b_j^+))$. For each such i_0 , consider $j(i_0) \equiv j \in \{1, \dots, l\}$ such that $1/\theta_{i_0} \in (G_{\mu_X}(a_{j-1}^-), G_{\mu_X}(b_j^+))$ and let \tilde{u} be a unit-norm eigenvector of \tilde{X} associated with the eigenvalue $\tilde{\lambda}_{[nc_{j-1}] + i_0}$. Then we have, as $n \rightarrow \infty$,*

(a)

$$|\langle \tilde{u}, \ker(\theta_{i_0} I_n - P_n) \rangle|^2 \xrightarrow{a.s.} \frac{-1}{\theta_{i_0}^2 G'_{\mu_X}(\rho)}$$

where ρ is the limit of $\tilde{\lambda}_{[nc_{j-1}] + i_0}$ given by Theorem 5.1;

(b)

$$\langle \tilde{u}, \oplus_{i \neq i_0} \ker(\theta_i I_n - P_n) \rangle \xrightarrow{a.s.} 0.$$

Proof. The result is obtained by following the approach taken in [4, pp. 514-516] for proving Theorem 2.2 and accounting for the possibly multi-valued nature of $G_{\mu_X}^{-1}(1/\theta)$. \square

Theorem 5.3 (Phase transition of eigenvector informativeness). *When $r = 1$, let the sole non-zero eigenvalue of P_n be denoted by θ . Consider $j \in \{1, \dots, \ell\}$ such that*

$$\frac{1}{\theta} \notin (G_{\mu_X}(a_{j-1}^-), G_{\mu_X}(b_j^+)), \quad \text{and} \quad G'_{\mu_X}(b_j^+) = -\infty \text{ and } G_{\mu_X}(a_{j-1}^-) = -\infty.$$

For each n , let \tilde{u} be a unit-norm eigenvector of \tilde{X} associated with $\tilde{\lambda}_{[nc_{j-1}] + 1}$. Then we have

$$\langle \tilde{u}, \ker(\theta I_n - P_n) \rangle \xrightarrow{a.s.} 0,$$

as $n \rightarrow \infty$.

Proof. The result is obtained by following the approach taken in [4, pp. 516-517] for proving Theorem 2.3 and accounting for the possibly multi-valued nature of $G_{\mu_X}^{-1}(1/\theta)$. \square

The following proposition allows to assert that in many classical matrix models, such as Wigner or Wishart matrices, the above phase transitions actually occur with a finite threshold.

Proposition 5.4 (Edge density decay condition and the phase transition). *Assume that the limiting eigenvalue distribution μ_X , supported on ℓ disjoint intervals, has a density f_{μ_X} with a power decay at b_j for $j = 1, \dots, \ell$, i.e., that, as $t \rightarrow b_j$ with $t < b_j$, $f_{\mu_X}(t) \sim c(b_j - t)^\alpha$ for some exponent $\alpha > -1$ and some constant c . Then:*

$$G_{\mu_X}(b_j^+) < \infty \iff \alpha > 0 \quad \text{and} \quad G'_{\mu_X}(b_j^+) = -\infty \iff \alpha \leq 1.$$

Similarly, if f_{μ_X} has a power decay at a_{j-1} for $j = 2, \dots, \ell$, i.e., that, as $t \rightarrow a_{j-1}$ with $t > a_{j-1}$, $f_{\mu_X}(t) \sim c(t - a_{j-1})^\alpha$ for some exponent $\alpha > -1$ and some constant c . Then

$$G_{\mu_X}(a_{j-1}^-) < \infty \iff \alpha > 0 \quad \text{and} \quad G'_{\mu_X}(a_{j-1}^-) = -\infty \iff \alpha \leq 1.$$

Theorem 5.1 describes the fundamental limits of eigen-gap based signal detection. Principal eigen-gap detection will fail whenever $\theta_i < 1/G_{\mu_X}(b_1^+)$. If $G_{\mu_X}(b_j^+) > G_{\mu_X}(b_1^+)$ for $j = 2, \dots, \ell$ then principal eigen-gap detection will be suboptimal as the middle eigen-gaps will reveal the presence of a low-rank signal even when the principal eigen-gap does not. Theorem 5.2 shows that whenever there is an eigen-gap, the corresponding eigenvectors will be informative. Theorem 5.3 provides insight on the fundamental limits of low-rank signal matrix estimation.

5.2. Singular values and singular vectors. Let X_n be an $n \times m$ ($n \leq m$, without loss of generality) random matrix whose ordered singular values we denote by $\sigma_1(X_n) \geq \dots \geq \sigma_n(X_n)$. Let μ_{X_n} be the empirical singular value distribution, i.e., the probability measure defined as

$$\mu_{X_n} = \frac{1}{n} \sum_{i=1}^n \delta_{\sigma_i(X_n)}.$$

As before, assume that the probability measure μ_{X_n} converges almost surely weakly, as $n \rightarrow \infty$, to a non-random compactly supported probability measure μ_X that is supported on ℓ disjoint intervals so that

$$\mu_X = \sum_{j=1}^{\ell} p_j \mu_j,$$

where for $j = 1, \dots, \ell$, the measures $\mu_j(x)$ is a non-random probability measure are supported on $[a_j, b_j]$ with $d\mu_j(z) > 0$ for $z \in (a_j, b_j)$ and $a_\ell < b_\ell < a_{\ell-1} < b_{\ell-1} < \dots < a_1 < b_1$. Define $c_j = \sum_{i=0}^j p_i$ with $p_0 := 0$, $c_0 := 0$ and $c_\ell := 1$. We assume that for $j = 1, \dots, \ell$, $\sigma_{\lceil nc_{j-1} \rceil + 1} \xrightarrow{\text{a.s.}} b_j$ and that $\sigma_{\lceil nc_j \rceil} \xrightarrow{\text{a.s.}} a_j$. As before, we use $\lceil nc_{j-1} \rceil$ to denote the smallest integer greater than or equal to nc_{j-1} .

For a given $r \geq 1$, let $\theta_1 \geq \dots \geq \theta_r$ be deterministic non-zero real numbers, chosen independently of n . For every n , let P_n be an $n \times m$ matrix having rank r with its r non-zero singular values equal to $\theta_1, \dots, \theta_r$. We suppose that X_n and P_n are independent and that X_n , the noise-only matrix is bi-unitarily invariant while the low-rank signal matrix P_n is deterministic. Recall that a random matrix is said to be *bi-orthogonally*

invariant (or *bi-unitarily invariant*) if its distribution is invariant under multiplication on the left and right by orthogonal (or unitary) matrices. Alternately, if P_n has isotropically random right (or left) singular vectors then, then X_n need not be unitarity invariant under multiplication on the right (or left, resp.) by orthogonal or unitary matrices. Equivalently, X_n can have deterministic right and left singular vectors while P_n can have isotropically random left and right singular vectors and we would get the same result stated shortly.

Consider the rank r additive perturbation of the random matrix X_n given by

$$\tilde{X}_n = X_n + P_n.$$

where

$$P_n = \sum_{i=1}^r \theta_i u_i v_i^*,$$

and $\{u_i\}_{i=1}^r$ and $\{v_i\}_{i=1}^r$ are the left and right singular vectors, respectively of P_n .

For this model, we establish the following results.

Theorem 5.5 (Largest singular value phase transition). *The singular values of \tilde{X} exhibit the following behavior as $n, m_n \rightarrow \infty$ and $n/m_n \rightarrow c$. We have that for each $1 \leq i \leq r$ and $1 \leq j \leq \ell$,*

$$\sigma_{[nc_{j-1}] + i}(\tilde{X}) \xrightarrow{a.s.} \begin{cases} D_{\mu_X, (b_j, a_{j-1})}^{-1}(1/\theta_i^2) & 1/D_{\mu_X}(b_j^+) < \theta_i^2 < 1/D_{\mu_X}(a_{j-1}^-), \\ b_j & \text{if } \theta_i^2 < 1/D_{\mu_X}(b_j^+), \\ a_{j-1} & \text{if } \theta_i^2 > 1/D_{\mu_X}(a_{j-1}^-), \end{cases}$$

where D_{μ_X} , the D -transform of μ_X defined by

$$D_{\mu_X}(z) := \left[\int \frac{z}{z^2 - t^2} d\mu_X(t) \right] \times \left[c \int \frac{z}{z^2 - t^2} d\mu_X(t) + \frac{1-c}{z} \right] \quad \text{for } z \notin \cup_{j=1}^{\ell} [a_j, b_j],$$

and $D_{\mu_X, (b_j, a_{j-1})}^{-1}(\cdot)$ will denote its functional inverse on (b_j, a_{j-1}) with $a_0 = +\infty$.

Proof. The result is obtained by following the approach taken in [5, pp. 127–129] for proving Theorem 2.9 and accounting for the possibly multi-valued nature of the $D_{\mu_X}^{-1}(\cdot)$. The key ingredient of the proof is the recognition that the non-zero, positive eigenvalues of

$$\begin{bmatrix} 0 & \tilde{X} \\ \tilde{X}^* & 0 \end{bmatrix} = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix} + \begin{bmatrix} 0 & \sum_{i=1}^r \theta_i u_i v_i^* \\ \sum_{i=1}^r \theta_i v_i u_i^* & 0 \end{bmatrix},$$

are precisely the singular values of \tilde{X} . Thus adopting the approach outlined in Section 2 while taking into account the structured rank $2r$ perturbation gives us the stated result. \square

Theorem 5.6 (Informativeness of singular vectors). *Assume throughout that $\theta > 0$ and let $D_{\mu_X}(a_0^-) = +\infty$. Consider $i_0 \in \{1, \dots, r\}$ such that $1/\theta_{i_0}^2 \in \bigcup_{j=1}^{\ell} (D_{\mu_X}(a_{j-1}^-), D_{\mu_X}(b_j^+))$. For each such i_0 , consider $j(i_0) \equiv j \in \{1, \dots, \ell\}$ such that $1/\theta_{i_0}^2 \in (D_{\mu_X}(a_{j-1}^-), D_{\mu_X}(b_j^+))$*

and let \tilde{u} and \tilde{v} be unit-norm left and right singular vectors of \tilde{X} associated with the singular value $\tilde{\sigma}_{\lceil nc_{j-1} \rceil + i_0}$. Then we have, as $n \rightarrow \infty$,

a)

$$|\langle \tilde{u}, \text{Span}\{u_i; \theta_i = \theta_{i_0}\} \rangle|^2 \xrightarrow{a.s.} \frac{-2\varphi_{\mu_X}(\rho)}{\theta_{i_0}^2 D'_{\mu_X}(\rho)}, \quad (13)$$

b)

$$|\langle \tilde{v}, \text{Span}\{v_i; \theta_i = \theta_{i_0}\} \rangle|^2 \xrightarrow{a.s.} \frac{-2\varphi_{\tilde{\mu}_X}(\rho)}{\theta_{i_0}^2 D'_{\mu_X}(\rho)}, \quad (14)$$

where ρ is the limit of $\tilde{\sigma}_{i_0}$ given by Theorem 5.5 and $\tilde{\mu}_X = c\mu_X + (1-c)\delta_0$ and for any probability measure μ ,

$$\varphi_\mu(z) := \int \frac{z}{z^2 - t^2} d\mu(t). \quad (15)$$

c) Furthermore, in the same asymptotic limit, we have

$$|\langle \tilde{u}, \text{Span}\{u_i; \theta_i \neq \theta_{i_0}\} \rangle|^2 \xrightarrow{a.s.} 0, \quad \text{and} \quad |\langle \tilde{v}, \text{Span}\{v_i; \theta_i \neq \theta_{i_0}\} \rangle|^2 \xrightarrow{a.s.} 0,$$

and

$$\langle \varphi_{\mu_X}(\rho) P_n \tilde{v} - \tilde{u}, \text{Span}\{u_i; \theta_i = \theta_{i_0}\} \rangle \xrightarrow{a.s.} 0.$$

Proof. The result is obtained by following the approach taken in [5, pp. 129–131] for proving Theorem 2.10 and accounting for the possibly multi-valued nature of the $D_{\mu_X}^{-1}(\cdot)$. \square

Theorem 5.7 (Phase transition of vector informativeness). *When $r = 1$, let the sole singular value of P_n be denoted by θ . Consider $j \in \{1, \dots, \ell\}$ such that*

$$\frac{1}{\theta^2} \notin (D_{\mu_X}(a_{j-1}^-), D_{\mu_X}(b_j^+)), \quad \text{and} \quad D'_{\mu_X}(b_j^+) = -\infty \text{ and } D'_{\mu_X}(a_{j-1}^-) = -\infty.$$

For each n , let \tilde{u} and \tilde{v} be unit-norm left and right singular vectors of \tilde{X} associated with $\tilde{\sigma}_{\lceil nc_{j-1} \rceil + 1}$. Then we have that

$$\langle \tilde{u}, \ker(\theta^2 I_n - P_n P_n^*) \rangle \xrightarrow{a.s.} 0, \quad \text{and} \quad \langle \tilde{v}, \ker(\theta^2 I_m - P_n^* P_n) \rangle \xrightarrow{a.s.} 0,$$

as $n \rightarrow \infty$.

Proof. The result is obtained by following the approach taken in [5, pp. 131] for proving Theorem 2.11 and accounting for the possibly multi-valued nature of the $D_{\mu_X}^{-1}(\cdot)$. \square

Theorem 5.1 describes the fundamental limits of eigen-gap based signal detection. Principal gap detection will fail whenever $\theta_i^2 < 1/D_{\mu_X}(b_1^+)$. If $D_{\mu_X}(b_j^+) > D_{\mu_X}(b_1^+)$ for $j = 2, \dots, \ell$ then principal eigen-gap detection will be suboptimal as the middle eigen-gaps will reveal the presence of a low-rank signal even when the principal eigen-gap does not. Theorem 5.6 shows that whenever there is an eigen-gap, the corresponding singular vectors will be informative. Theorem 5.7 provides insight on the fundamental limits of low-rank signal matrix estimation. The analog of Proposition 5.4 also applies here.

6. NOISE MODELS THAT MIGHT PRODUCE INFORMATIVE MIDDLE COMPONENTS

Our discussion has brought into sharp focus the pivotal role played by the noise eigen-spectrum in determining the relative informativeness of the principal and middle components of the singular value (or eigen) decomposition of signal-plus-noise data matrix models as in (1).

Specifically, we showed that if the noise eigen-spectrum is supported on a single connected interval then the principal components will indeed (with high probability) be the most informative components and their use in detection and estimation is justified.

However, if the noise eigen-spectrum is supported on multiple intervals, as in Figure 8, then the principal components will remain informative in the high SNR regime (*i.e.*, large θ_i). However, for moderate to low SNR, the middle components might also be informative and may remain informative even when the principal components are no longer informative. In such settings, identifying large middle eigen-gaps and using the associated middle eigenvectors for inference can improve inference.

This leads to a natural question: When will the noise eigen-spectrum exhibit a disconnected spectrum?

We conclude by identifying a large class of Gaussian mixture models that produce precisely such an eigen-spectrum. Consider the class of noise matrices modeled as

$$X = G\Sigma^{1/2},$$

where G is an $m \times n$ matrix with i.i.d. mean zero, variance $1/m$ (say) Gaussian entries. If the rows of X denote spatial measurements and the columns represent temporal measurements, then Σ is a temporal covariance matrix and XX^* is a Wishart distributed matrix. These models arise in many statistical signal processing and machine learning applications where PCA/SVD is often used as the first step in inferential process (see, for e.g. [32, 7, 28, 33, 20, 16]).

Bai and Silverstein characterize the limiting eigenvalue distribution of XX^* in [29]. What emerges from their analysis [1, 31] is that for the noise eigen-spectrum of XX^* to have a disconnected spectrum, the eigenvalue spectrum of Σ has to have a limiting distribution that is supported on disconnected intervals. In addition, the separation between these intervals has to be relatively large for the spectrum of X to be supported on disconnected intervals. There is no simple formula for how large this separation has to be. There are expressions in [29] for the form of the spectrum of X as a function of the spectrum Σ , from which it can be ascertained whether the support is supported on multiple intervals or not using the results in [1, 31]. Moreover, the spectrum will exhibit square-root decay at the edges [30] and so the phase transitions described will manifest.

Figure 9 plots evolution of the i -th singular value and $\text{Inf}_i = |\langle \tilde{u}_i, u \rangle|^2$ as a function of θ for the model in (1) with $X = G\Sigma^{1/2}$ and $\Sigma = \text{diag}(20I_{n/10}, I_{n-n/10})$. The figure clearly shows the phase transition in the informativeness of the principal and middle components and shows that there is a low SNR regime where the middle component is informative even when the principal component is not. The values where the phase transitions occur can be theoretically predicted, if so desired, using Theorem 5.5. Figure 2(a) shows a sample realization of the singular values for the same setting when $\theta = 2$.

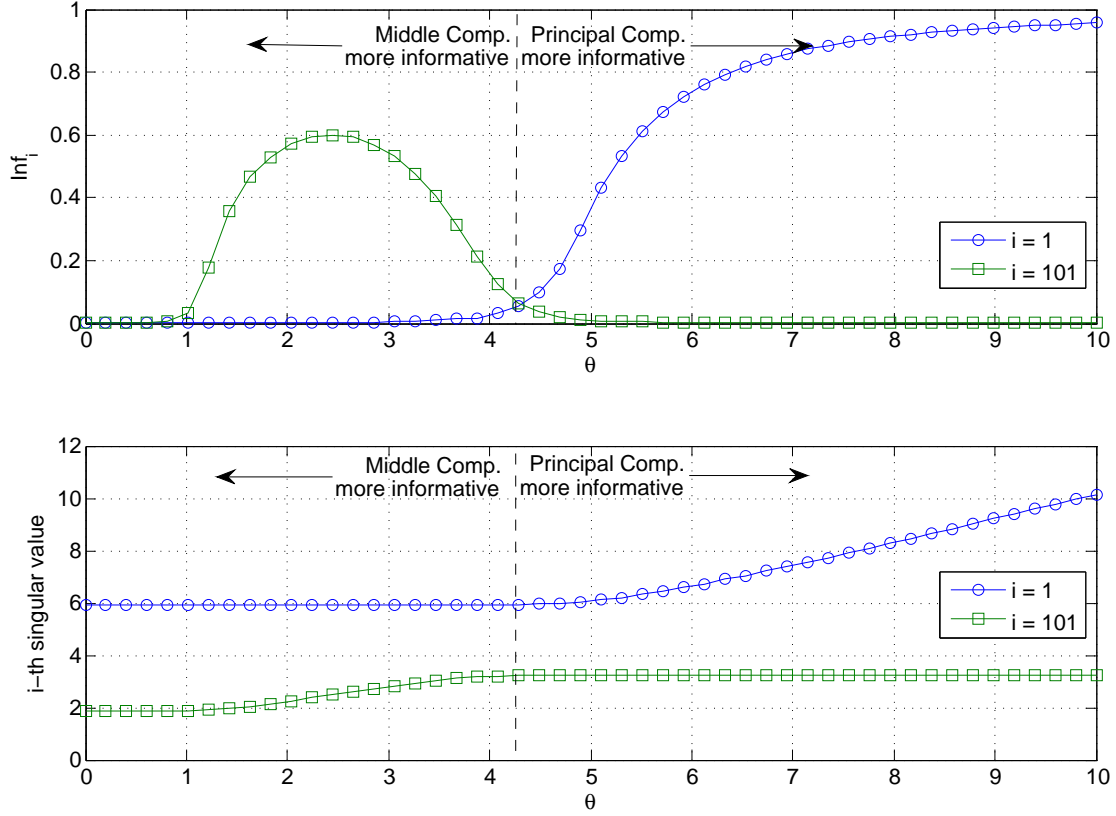


FIGURE 9. The evolution of the informativeness of the principal and the middle component as a function of θ for the model in (1) and $X = G\Sigma^{1/2}$, where G is an $n \times m$ matrix (here $n = m = 1000$) with i.i.d. mean zero, variance $1/m$ entries and $\Sigma = \text{diag}(20I_{n/10}, I_{n-n/10})$. The upper panel plots $\text{Inf}_i = |\langle \tilde{u}_i, u \rangle|^2$ computed over 250 Monte-Carlo trials. The lower panel plots the i -th largest singular value of \tilde{X} .

Thus the potential for informative middle components to emerge is the greatest in large, heterogeneous datasets where there might be significant temporal (or spatial) variation. These might be exploited for extracting additional processing gain beyond what principal component analysis might offer.

Conversely, if the temporal covariance matrix Σ represents a relatively homogenous (in time) data set, then there will be no gap in the eigen-spectrum and the use of principal components is justifiably optimal.

Expanding the range of noise models for which similar predictions can be made is a natural next step. It remains an open problem to fully characterize the vanishing informativeness of the components of the singular value decomposition associated with singular values that (asymptotically) exhibit an $o(1)$ eigen-gap. Additional hypotheses on the noise eigen-distribution will likely be required - establishing natural conditions for these remains an important line of inquiry. A result along these lines would firmly establish that the informative components associated with the singular/eigen values that exhibit an eigen-gap are indeed the *maximally* informative components.

REFERENCES

- [1] ZD Bai and J.W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.
- [2] J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.
- [3] J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.
- [4] F. Benaych-Georges and R.R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.
- [5] F. Benaych-Georges and R.R. Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, pages 120–135, 2012.
- [6] M. Brand. Fast low-rank modifications of the thin singular value decomposition. *Linear algebra and its applications*, 415(1):20–30, 2006.
- [7] S. Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.
- [8] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 292–303, 2006.
- [9] P. Drineas and M.W. Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *The Journal of Machine Learning Research*, 6:2153–2175, 2005.
- [10] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [11] N. El Karoui. Tracy–widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, 2007.
- [12] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2001.
- [13] A. Frieze, R. Kannan, and S. Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.
- [14] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [15] Fumio Hiai and Dénes Petz. *The semicircle law, free random variables and entropy*, volume 77 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2000.
- [16] D. Hsu and S.M. Kakade. Learning gaussian mixture models: Moment methods and spectral decompositions. *arXiv preprint arXiv:1206.5766*, 2012.
- [17] I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis.(english. *Ann. Statist.*, 29(2):295–327, 2001.
- [18] I.M. Johnstone. High dimensional statistical inference and random matrices. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 307–333, 2006.
- [19] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.
- [20] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *Learning Theory*, pages 155–199, 2005.
- [21] S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.
- [22] S. Kritchman and B. Nadler. Non-parametric detection of the number of signals: hypothesis testing and random matrix theory. *Signal Processing, IEEE Transactions on*, 57(10):3930–3941, 2009.
- [23] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.
- [24] R.R. Nadakuditi and A. Edelman. Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *Signal Processing, IEEE Transactions on*, 56(7):2625–2638, 2008.
- [25] B. Nadler. Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator. *Signal Processing, IEEE Transactions on*, 58(5):2746–2756, 2010.

- [26] A. Onatski. Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics*, 92(4):1004–1016, 2010.
- [27] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.
- [28] A. Sanjeev and R. Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.
- [29] J.W. Silverstein and ZD Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995.
- [30] J.W. Silverstein and S.I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.
- [31] J.W. Silverstein and P.L. Combettes. Signal detection via spectral theory of large dimensional random matrices. *Signal Processing, IEEE Transactions on*, 40(8):2100–2105, 1992.
- [32] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.
- [33] S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 113–122. IEEE, 2002.

RAJ RAO NADAKUDITI, DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE,
UNIVERSITY OF MICHIGAN, 1301 BEAL AVENUE, ANN ARBOR, MI 48109. USA.

E-mail address: `rajnrao@eecs.umich.edu`

URL: `http://www.eecs.umich.edu/~rajnrao/`